情报学报　2018 年 3 月　第 37 卷　第 3 期
Journal of the China Society for Scientific and Technical Information, Mar. 2018, 37(3): 318-328

DOI: 10.3772/j.issn.1000-0135.2018.03.009

科学计量中多源数据融合方法研究述评

许海云[1,2]，董　坤[1,3]，隗　玲[1,3]，王　超[1,3]，岳增慧[4]

（1. 中国科学院成都文献情报中心，成都　610041；2. 中国科学技术信息研究所，北京　100038；
3. 中国科学院大学，北京　100190；4. 济宁医学院医学信息工程学院，日照　276826）

摘　要　本文系统综述了科学计量学中多源数据融合的研究和应用现状，将科学计量中多源数据融合划分为前期融合、中期融合和后期融合，并重点分析了多数据类型关系的获取与融合实现方法。在此基础上，提出科学计量分析中多源数据融合的挑战和未来可能的突破方向，并借鉴数理领域的数据融合方法，构建了未来科学计量分析的多源数据融合流程和发展趋势。

关键词　多源数据；多元关系融合；数据融合；多源异构网络；科学计量

Research on Multi-source Data Fusion Method in Scientometrics

Xu Haiyun[1,2], Dong Kun[1,3], Wei Ling[1,3], Wang Chao[1,3] and Yue Zenghui[4]

(1. *Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu*　610041;
2. *Institute of Scientific and Technical Information of China (ISTIC), Beijing*　100038; 3. *University of Chinese Academy of Sciences, Beijing*　100190; 4. *School of Medical Information Engineering, Jining Medical University, Rizhao*　276826)

Abstract: This paper systematically reviews the research and application of multi-source data fusion in Scientometrics. Multi-source data fusion in scientific metrology can be divided into early fusion, mid-term fusion, and post-fusion, and focuses on the relationship between multi-data type acquisition and the fusion of multiple data type relationships. Subsequently, the challenge of multi-source data fusion in scientific econometric analysis and the possible future breakthrough are put forward. Based on the fusion method in mathematics, the multi-source data fusion process and trend of development of future scientific econometric analysis are constructed.

Key words: multi-source data; multi-relations fusion; data fusion; multi-source heterogeneous network; scientometrics

1　引　言

科学计量学诞生于 20 世纪 60 年代初，是一门用数学方法研究科学的量化特征和发展机制的学科。科学计量学通过对大量有价值的信息的有效、定量分析，获取与存储、知识的生产与流动、知识的离散与重组，辅助学科领域的研究者发现领域研究的热点及前沿，并作出科学评价[1]。科学计量学在经过多年发展后，当前面临的最大问题依旧是在评价和决策支撑中常常出现偏差、无法脱离专家意见

收稿日期：2017-11-01；修回日期：2017-12-11
基金项目：国家自然科学基金项目“基于科学-技术主题关联分析的创新演化路径识别方法研究”（71704170）；中国博士后科学基金资助项目“面向多关系融合的知识创新路径的识别与预测方法研究”（2016M590124）；中国科学院青年创新促进会（2016159）。
作者简介：许海云，女，1982 年生，博士，副研究员，研究方向为情报计量学的理论与实践，E-mail: xuhy@clas.ac.cn；董坤，女，1990 年生，博士研究生，研究方向为专利情报分析；隗玲，女，1981 年生，博士研究生，研究方向为情报计量学的理论与实践；王超，男，1988 年生，博士研究生，研究方向为竞争情报分析；岳增慧，女，1985 年生，博士，大学讲师，研究方向为情报计量学的理论与实践。

指导的局限，因而难以发挥更大的作用。深究其原因在于当前的大多科学计量方法往往采用单一数据源与数据关系作为分析基础，导致分析广度和深度无法与专家知识相比。

Morris 等[2]指出基于某种关系的科学计量方法只能从某一方面反映对科学领域的有限认识，即每一种关系只能有助于研究人员从某个特定的角度去分析研究领域的局部特征。只有从不同的角度观察某个研究领域，才有可能全面地了解它，整合多重关系有助于对问题的全面理解。同时，当前伴随着科技文献数量的爆炸式增长以及文献类型的不断丰富，科学计量可分析的关系类型也在不断地拓展，除了传统的引用和共现关系外，不同实体间的多种耦合关系分析也被成功运用，多源异构成为重要且常见的数据存在形式。此种情况下，如何通过数据融合处理好多源异构数据成为科学计量面临的新问题[3-4]，也是科学计量学发展的新契机。如何充分利用当前多类型的数据和多种计量关系，将科学计量方法改进为多源数据融合下的信息分析方法，使得科学计量分析通过多源数据融合的方式，接近甚至超越专家分析的准确度，是科学计量分析能力提高的重要突破方向。

本文将在系统综述科学计量分析中多源数据融合研究和应用现状的基础上，总结科学计量中多源数据融合面临的问题和挑战，并结合自动化、生物健康等领域的数据融合最新进展。在此基础上，针对科学计量的数据分析特点，拓展科学计量中的多源数据融合方法。

2　科学计量的多源数据融合分析

2.1　多数据源融合的概念

多源数据融合（以下简称数据融合）是指通过特定的方法对不同类型信息来源或关系数据进行综合分析，最终可以利用所有信息共同揭示研究对象的特征，弥补单一数据类型与单一关系类型在揭示研究领域实体间关联的不足，以获取更全面、客观的计量结果。

化柏林[5]提出了将融合论作为情报学研究的主要方法论之一，并强调了数据融合、信息融合和知识融合在情报学中的重要作用，随后，他进一步论述了数据融合对情报工作的重要性，分析了不同类型多源信息的融合方法[6]。此外，化柏林等[7]还从多源信息融合的表示形式流程、技术算法与模型的角度，系统阐述了信息融合的相关理论和应用。

2.2　科学计量中多数据源融合方法

科学计量中多源数据融合可以按照融合阶段不同划分为三种类型：前期融合、中期融合和后期融合。其中，前期融合是为数据源与数据类型的融合，中期融合为数据关系融合，后期融合为聚类融合。

1）前期融合——数据源融合

前期融合包括将不同的数据源融入同一分析目标。当前数据源主要有期刊论文、会议信息、学位论文、专利信息、项目信息、著作信息以及科研履历等。化柏林[6]把多源信息划分为同质异源信息、异质异构信息以及多语种信息，并指出数据类型的融合是数据融合中的一项基础性工作，涉及字段映射、字段拆分、数据记录滤重、异构数据加权，是未来科学计量分析中数据处理的必然步骤。曲建升等[8]研究了基于文献的知识发现的研究背景及现有异构信息标准化的方式，提出处理异构信息需要解决数据的多元性及多源性问题，以数据和信息集成过程中异构信息的标准化为研究目标，提出了异构信息标准化处理的思路。本文重在综述多种数据类型之间关系融合，关于多种数据源融合方法不做过多介绍。

数据源中的实体类型包括科技文献本身、作者、机构、主题词、期刊、参考文献以及参考文献的作者、所在期刊等。以科技文献为代表的不同实体之间的关联关系是科学计量学的重要研究内容。引文链接、词间关系和作者合著关系是最主要的关系类型。实体类型之间根据数据关系类型的不同，引用关系和词间关系不仅可以用于描述文献间的关系，也可以扩展至作者、期刊、机构间的关系描述和分析。图 1 展示了当前科学计量中的主要数据源与分析实体类型。第三层为主要数据源的载体类型，第二层为科学计量的主要数据源，第一层为计量的实体类型。多种实体及其关系类型构成科学计量的多源异构网络。

2）中期融合——数据关系融合

数据关系融合是指获取多种数据关系，对不同数据关系有效融合成一个新关系数据来表征实体之间的关联特征，新的关系将多源数据的相似矩阵或距离矩阵整合为一个新的关系数据。多元关系也常被称为多模关系、多维关系、多重关系等。如图 2 所示，通过融合共现、引用和合著关系可以形成计量目标的多元关系网络。

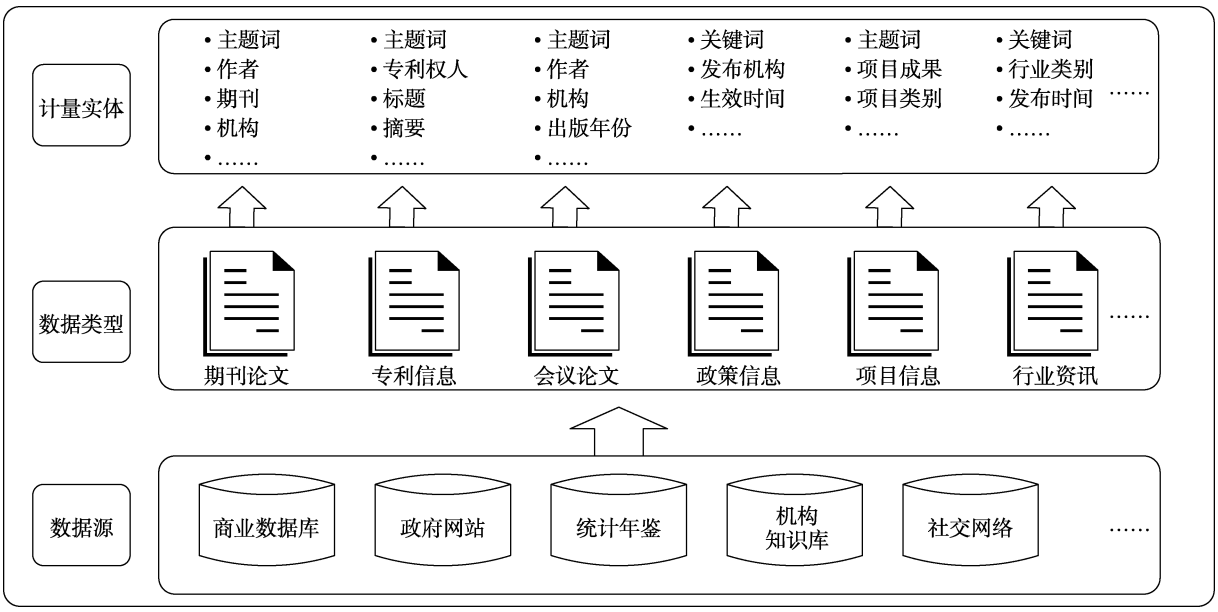

图 1 科学计量中数据源与数据类型的融合示意图

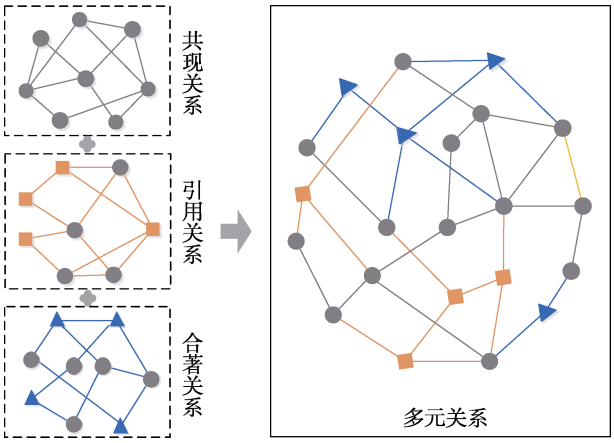

图 2 数据关系融合示意图

当前科学计量已有不少数据关系融合研究与实践。Small[9]采用引用链接和共词两种方法发现了文献间的直接连接和间接连接关系。Glenisson 等[10]结合全文本主题分析与文献计量分析，通过大规模数据集证明了方法的可行性。van Den Besselaar 等[11]提出利用主题词与参考文献共现方式描述研究主题，并将利用这种共现关系聚类而形成的文献集称为“研究前沿”。Jo 等[12]分析了引文图中的词分布和链接分布的相关性，并发现引文链接与文本主题词之间相关性科学计量中多种实体关系的融合常被称为多模变量融合分析。Shibata 等[13]通过特征词聚类的演化发现相似的主题通过交叉引用聚集在一起，而不同的主题之间链接很弱。Wen 等[14]分别得到基于期刊互引网络、作者关键词共现网络与题名主题词-引文共现网络三种网络的主题结构，利用交叉对照表形成三种主题的交叉验证，实证证明三种方法可以在主体结构识别中结合运用，并能更好的识别跨学科领域。Dong 等[15]构建了基于科技文献内容分析的多维学科交叉主题识别方法及流程，旨在通过集成不同维度的分析结果达到相互印证与补充的效果，实证表明该方法及流程在识别热点学科交叉主题的同时还能有效识别潜在学科交叉主题，可获得更加全面、准确的学科交叉主题识别结果。

多元关系融合根据不同的分类依据存在多种不同的分类方式。根据关系融合的广度与深度，可以分为同种对象关系融合与不同对象关系融合；根据融合的具体方式的不同，可以分为多模数据交叉融合和多关系矩阵融合；根据融合关系的数量，可以分为两种关系融合与多种关系融合；根据融合关系是过滤性还是兼容性，可以分为串行融合与并行融合。其中，串行融合是指在一种关系下再考虑其他关系，实质上是用一种关系来限制另一种关系，并行融合是指同时考虑多种关系。

3）后期融合——聚类融合

多元关系融合之后需要对最终的综合型矩阵进行聚类分析，不同的聚类算法往往会得到不一致的聚类结果。如何对多种结果合理取舍或整合各种不同的结果，得到更合理的聚类，需要聚类集成分析。聚类集成即将不同关系数据先分别进行聚类，然后将不同聚类结果通过融合函数合并为一个聚类结果。

聚类融合是将不同算法或者同一算法使用不同参数得到的大量聚类成员利用融合函数进行融合，

从而获得最终聚类结果[16]。它的具体表达如下：给定包含 N 个数据对象的数据集 $X=\{x_1,x_2\cdots x_N\}$，对数据集 X 进行 H 次聚类得到 H 个聚类结果的聚类成员集 $P=\{p_1,p_2\cdots p_H\}$，其中第 i 次聚类结果 $P_i=\{C_i^1,C_i^2,\cdots C_i^{k_i}\}$，$(i=1,2,\cdots,H)$，$k_i$ 是聚类成员 P_i 的簇数。聚类融合的目的就是设计一种融合函数 Γ，将所有的聚类成员 $p_1,p_2\cdots p_H$ 合并得到最终的聚类结果 P^f。聚类融合的示意图如图 3 所示。

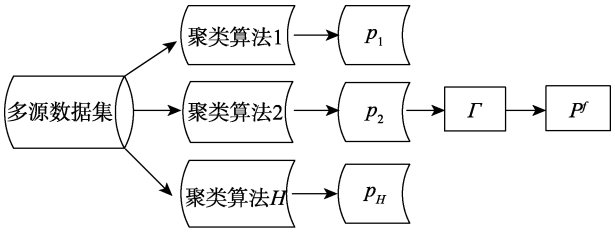

图 3　聚类融合示意图

从图 3 可以看出，聚类融合算法分为两个过程，首先分别利用不同的聚类算法产生大量的初始聚类成员，之后将初始聚类成员利用融合函数组合得到最终聚类结果。与单一的聚类算法相比，聚类融合通过从不同侧面的数据集的特征组合提高了聚类的性能。同时，聚类融合适于并行处理数据，特别是分布式的数据集，可以对每个分布的数据集进行并行的聚类，然后将得到的多个聚类结果融合得到最终的聚类结果。融合函数又称共识函数、一致性函数，是聚类集成需要重点解决的问题。常用的融合函数包括基于 Co-association 矩阵方法、超图划分的方法、基于信息论的方法、基于混合模型的方法、基于投票的方法、基于证据累积的方法、基于神经网络的方法等。

从当前科学计量中数据关系融合研究与实践看出，三种数据融合类型各有侧重。前期融合侧重数据的和信息集成过程中异构信息的标准化，是科学计量数据融合中的基础性工作。中期融合是数据融合的关键，在不同的融合任务中，各种关系类型承担的作用不同，因此，如何处理不同的关系类型是关系融合的研究重点。后期融合方法通过最优的共识函数实现聚类融合，可以进一步优化多源数据融合，目前在科学计量分析中应用还不充分。

3　多数据类型关系与抽取方法

3.1　科技文献中的多元关系类型

Morris 等[70]对科技文献中常见的计量实体进行了概述，主要包括科技文献本身（以下简称为“目标文献”）、参考文献、科技文献所在期刊、科技文献作者、参考文献所在期刊、参考文献作者、主题词以及科技文献作者所在机构等。在由计量实体间关联而成的复杂网络中，任意两类计量实体以及同一类计量实体的任意两个个体都可以通过一定的连接路径产生关联；根据路径的长短可分为直接关联和间接关联；根据关联个体的类型又可分为内部关联和外部关联；根据信息源的远近带来的信息有效性不同，目标文献的重要性最高，其他文献随距离目标文献的远近而变化，距离越远重要性越低。本文结合主要计量实体的关联关系示意图对多种关联关系进行详细阐述。图 4 为常用计量实体的关联关系的示意图。

（1）外部直接关联关系。假设每种类型的计量实体均由 n 个个体组成，那么不同类型的计量实体（任意个体）之间，若能够通过单一路径发生直接关联，则认为这些计量实体之间存在直接关联关系。如图 4a 所示，同一篇目标文献对应的作者、主题词、来源期刊、引证文献与参考文献中，任意两类计量实体之间均为直接关联（用实线箭头表示）。

（2）外部间接关联关系。假设每种类型的计量实体均由 n 个个体组成，那么不同类型的计量实体（任意个体）之间，若不能通过单一路径发生直接关联，则认为两者存在间接关联。如图 4b 所示，目标文献 p 的作者 a 与参考文献 r 的作者 r_a 需要经过目标文献与参考文献间的双重路径产生联系，则认为 a 与 r_a 之间是一种间接关联（用虚线箭头表示）。

（3）内部直接关联关系。假设每种类型的计量实体均由 n 个个体组成，那么对于同一类型的两个个体而言，如果都可以通过单一路径与某一其他类型计量实体相关联，则认为两者存在直接关联关系。如图 4c 所示，一篇目标文献的作者包括 a_1 与 a_2，那么 a_1 与 a_2 都可以通过单一路径与目标文献关联，因此 a_1 与 a_2 具有直接关联（用实线箭头表示）。

（4）内部间接关联关系。假设每种类型的计量实体均由 n 个个体组成，那么对于任意类型的两个个体而言，若两者分别直接关联的其他计量实体不能通过单一路径相关联，则认为这两者之间是一种间接关联关系。如图 4d 所示，一个目标文献集由 n 篇目标文献组成，那么文献 p_1 的作者 a_1 与文献 p_2 的作者 a_2 之间不能通过单一路径相连，因此 a_1 与 a_2 具有间接关联（用虚线箭头表示）。

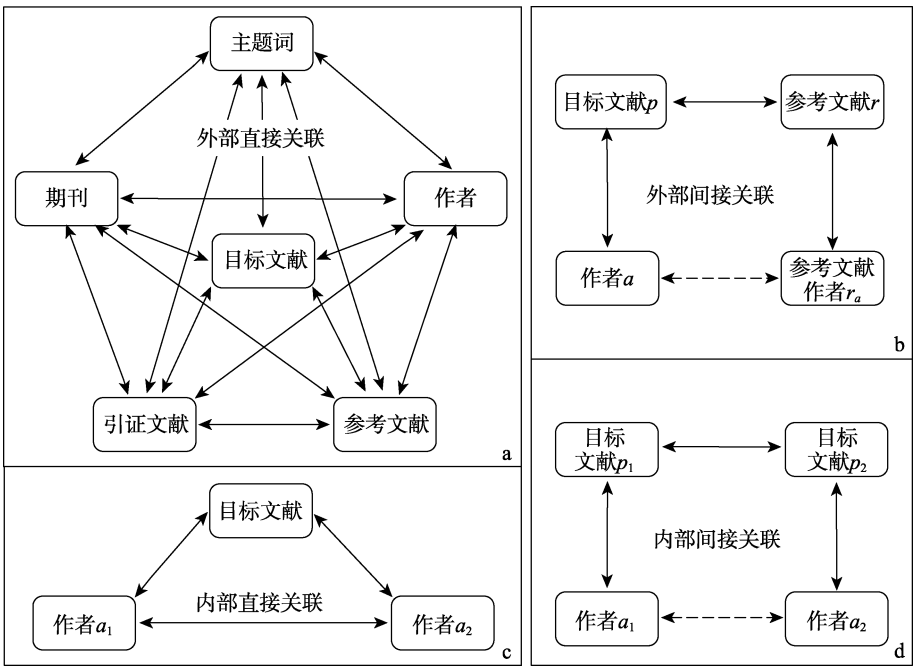

图 4 计量实体的关联关系示意图

3.2 基于元路径的多关系表示与抽取

元路径是链接对象间关系的组合。元路径能从不同角度呈现对象之间的相似性，富含语义[17]。在异构信息网络中，不同对象之间可以通过不同路径进行链接，且不同路径表示对象之间的不同关系，这些路径被称为元路径。元路径是指给定一个异构信息网络 $G(A,R)$，A 的元路径 P 定义为：$A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_L} A_{L+1}$，表示不同节点对象 A_1 和 A_{L+1} 之间的关系组合，示例如图 5 所示。图中双向箭线直接连接的任意两个不同类型的对象之间都组成了元路径，任意两个对象也可以借助中介对象和中介关系组成元路径。

元路径包含丰富的语义信息，不同对象之间可以通过不同链接路径表征不同关系。由此，不同元路径可以呈现主题词节点和不同关联属性组合形成的特定语义，也可以计算特定语义下主题词节点之间的相似度。目前基于元路径提取进行异构信息网络融合研究不断涌现。PathSim[17]算法成功应用于异构信息网络聚类研究中，结果显示基于元路径提取的聚类结果要优于仅依赖网络结构本身而进行的聚类。Sebastian[18]指出文献计量的异构网络提供了识别科技论文之间关联的潜在的和半结构化的文献计量信息，他利用从异构网络建立元路径的方法形成新的基于文献的知识发现方法，实现预测未来的同被引链接。Jensen 等[19]为了更好地表示和量化多类实体以及实体间的关系，采用元路径方法将文献、主题词、作者、引文等属性关联起来，从不同角度呈现对象之间的关联性和相似性，并初步应用于主题演化探索。

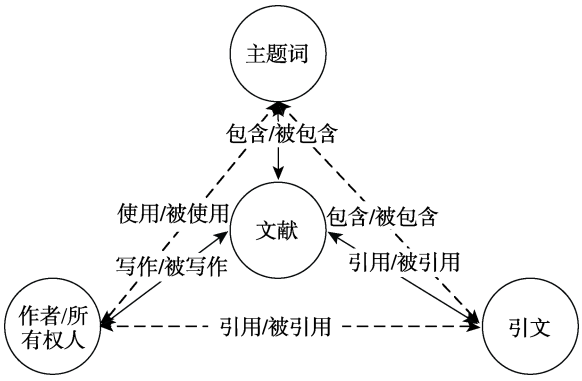

图 5 元路径示意图

多数据类型关系由于计量实体间关联而成复杂网络，因此，如何抽取出分析目标中的关键实体关联是实现数据融合的关键点。本文通过关联个体的直接、间接关联以及内部、外部关联，构建了科学计量实体的多重关联关系。同时，元路径不但可以从共现的视角决定所选元路径的长度，还能从语义的角度考虑不同节点对象之间的关系，提供丰富的语义特性。因此，元路径分析成为多元数据抽取的可行方法，而构建长度不同的元路径并计算主题词基于多种元路径的相似度是元路径分析的重点。

4　多数据类型关系的融合方法与应用

在获取了多数据类型关系之后，需要融合多数据关系。当前主要的多数据类型关系融合方法主要有五种：投影融合方法、非投影融合方法、超网络融合方法、非负矩阵关系融合方法和多模数据可视化融合方法。其中投影融合方法是较为简单和成熟的方法，其余四种随着近几年复杂网络分析、机器学习算法和可视化技术的突破，正在逐步引入科学计量分析中。

4.1　投影融合方法

投影法通过投影的方式把多模网络转化为 1 模网络，也称为单顶点网络或单分网络，然后进行网络分析。这种转化会导致信息的缺失而无法涵盖所有原始网络的性质。为保证分析过程的准确性，更可靠的方法是采用非投影融合法，直接分析原始多源异构网络的多种计量实体间的关系。

1）无加权投影融合方法

无加权投影不考虑各边的权重，是一种较为简便的处理方式。以作者-论文 2 模网络为例，将其转化为一模作者合作网络时，不考虑作者合作论文的数量，即合作强度，合作 1 次的作者与合作 10 次的作者之间的关系相同。Leydesdorff 等[20]在跨学科指标研究中，通过将非对称的 2 模互引矩阵与自身转置相乘，分别得到对称的 1 模同被引矩阵和耦合矩阵。无加权投影是一种比较经典的方法，其优势在于简单直观，将复杂的多模问题转化为简单的单模问题。但其局限性也显而易见：投影后造成重要信息的缺失，可能会使得整个网络边数剧增，可能会增加原网络中原本不存在的新信息[21]。现有的大量文献利用了异质学术网络中论文、作者和刊物这三类学术对象间的关系，但是现存研究都把这些复杂的多元多维关系投影成若干对二元关系或二元二维关系，这种扁平化的投影造成大量原始信息的丢失[22]。

2）加权投影融合方法

与无权投影不同的是加权投影给边增加了权属性，可以将信息压缩；更好地解释每组节点内容之间的关系；一定程度上消除信息无加权投影带来的信息损失。

目前存在多种不同加权投影方法的区别在于对权重的设定不同。最常见的一种方式是把权重定为两个同类节点共同连接的另一类节点的个数，如 Newman[23]提出基于作者合作论文数以及合作论文作者数目的加权网络，并将网络投影为一模的作者合作加权网络，再对网络进行聚类和中心性等分析。此外，Morris 等[24]提出了基于期刊论文构建 2 模加权网络的方法。设计了包括论文、论文作者、主题词、参考文献、参考文献作者等七种实体的加权耦合网络，并运用基于矩阵的权重函数计算带权二部矩阵和加权共现网络的实体关系权重。Evans 等[25]将加权线形图和 2 模网络投影的方法相结合提出一种重叠社区发现算法。Batagelj 等[26]在研究科研合作的过程中，利用矩阵乘法将 2 模网络转化成 1 模的合作网络，同时引入加权函数，该过程可视为一种加权的投影方法。Cerinšek 等[27]在研究科研合作及研究主题演化中，将科研成果网络表示成 4 个基本的 2 模网络：论文-作者网络、论文-期刊网络、论文-关键词网络和论文-主题分类网络，利用加权函数和矩阵乘法对各 2 模网络进行矩阵运算，最终得到复合多种信息的 1 模网络。

加权投影仍然无法涵盖原始网络的特性，且其权值的设定往往有很大的灵活性，社团划分的效果在很大程度上取决于权值。

4.2　非投影融合方法

把多模网络投影到 1 模网络上进行分析是一种比较经典的方法，但是无论何种投影方式都不能把两类节点的性质全部覆盖，因此，基于原始多模网络的直接分析逐渐开展起来。

周杰等[28]将对应分析方法与复杂网络相结合，提出利用改进的对应分析多元统计方法，进行学术研究主体之间、研究内容之间以及学术研究主体和研究内容之间的关联度计算，并采用 2 模网络图进行关联关系的可视化表示，实证发现该过程同时考虑同质关系和异质关系，避免了信息的损失。超网络构建是近几年兴起的多关系异质网络的非投影方法。Song 等[29]在研究化学-基因构成的异质网络的主题特征将 Newman 的模块化方法与 Barber 的 2 模网络划分方法通过线性加和。

Raghavan 等[30]介绍了一种标签传播算法用于社区挖掘，该算法首先将每个顶点分配一个唯一标签，然后在每次迭代过程中，每个顶点尽可能选取它们邻接顶点的标签，最后联系紧密且具有相同标签的顶点形成一个单独的社区。Latapy 等[21]将目前最为常用大型的 1 模网络统计指标引入 2 模网络特征分析，包括平均距离、度数、聚类系数、重叠度、冗余度等，并对比了指标在随机 2 模网络和真实 2 模

网络的度量差别，以寻求通过这些指标体系揭示 2 模网络的特征。Murata[32]在 Newman[31]的 1 模量化方法模块度的基础上，提出一种针对 2 模网络进行社区划分的算法。Barber[32]针对 2 模网络提出了模块化和社区识别算法，算法的合理思想是充分利用网络的两个模块之间存在相关性，即模块中的节点受到其他模块的影响。Comar 等[33]针对多关系异质网络提出了一种基于联合目标函数的多任务社区发现方法，网络中不但包括二分关系，还包括内部关系，该方法需要同时考虑链接结构、节点属性和网络中其他社区的属性。Liu 等[34]提出多关系异质网络模块划分算法，通过将异质网络拆解成 1 模网络、2 模网络和 3 模网络，并结合多种方法，以组合最优化为目标，该方法可以扩展到多种结构的网络中。

与投影法相比，非投影法的优势显而易见，更大限度地保持了原有网络信息的真实性和准确性，提高了社区划分的可靠性，但计算复杂度一般比转化为单模网络的投影法高出许多。

4.3 超网络融合方法

最早明确提出“超网络”概念的是美国科学家 Nagurney 等[35]。超网络是“高于而又超于现存网络的网络”，或者说“超网络”由是多个网络组成的网络[36]。在普通图中通过在两个节点之间的连接只能表示一对节点之间的关系，而在超图中的“边”，则可以包含任意多个节点，用来表示多个节点之间的关系。因此，基于超图的超网络动态演化模型则能够很好地描述和表示超网络中各节点之间的相互作用和影响，而只有对实际超网络的结构特征有很好的了解，才能建立合适的超网络模型。

目前，一个主要目标是将复杂网络的基于简单图的特征量的定义扩展到基于超图结构的超网络中。对于这些网络的一种自然的表示方法是用图的一种扩展图即为超图来表示[37]。滕立[38]基于超网络理论提出了构建作者-机构-国家混合共现网络，通过实证研究表明了该方法不但能够消除作者共现网络中的孤立节点，还可以通过附加作者与机构和国家间的耦合关系，丰富了混合网络的信息含量，使得总结领域中知识交流模式更加容易。高晨晖等[22]提出用有向同质超网建模文献之间和作者之间的多元引用关系，用异质超图建模文献-作者之间的多元多维关系，并使用异质超图来建模不同学术对象之间的无向多元多维关系，如作者和文献间的著作关系，以此将科学文献数据库建模为全面包含学术对象间多元多维关系的异质学术超网。

4.4 非负矩阵关系融合方法

Lange 等[39]提出使用非负矩阵因子分解方法进行数据融合，其本质为对从多信息源获得的相似矩阵进行组合。类似的，Wang 等[40]先对输入矩阵进行填充，然后对其进行分解。由于非负矩阵分解问题一般将原始矩阵分解为两个，因此，故该类分解被统称为二因子非负矩阵分解。Nickel 等[41]进一步提出三因子非负矩阵分解，将原始非负矩阵分解为三个矩阵，可以同时得到原始矩阵行和列的簇指示矩阵，并在簇指示矩阵上添加正交约束。新增的第三个矩阵为矩阵的分解提供了自由度，以保证因子矩阵很好的近似逼近原始矩阵。Wang 等于 2008 年和 2011 年分别提出 tri-SPMF（Symmetric penalized matrix factorization）[42]和 S-NMTF（Symmetric nonnegative matrix factorization）[43]，通过对称非负矩阵三因子分解对多类型关系数据进行聚类。Liu 等[44]提出了非负矩阵分解法对 2 模网络进行聚类划分，考虑 2 模节点的内部类型信息，揭示网络中隐含的结构信息，将具有收敛速度快、稳定性强等特点。Dunlavy 等[45]尝试利用多线性代数的张量分解方法，融合主题（包括关键词、题名和摘要三种相似度），作者和引文相似度。

此外，近年来有更多的数据融合方法出现在其他信息决策领域。Snidaro 等[46]对基于文本融合的信息系统做了系统介绍，该信息系统可以追述具体“概念”的想法来源。Xu 等[47]从信息融合的角度介绍了模糊决策理论和方法，主要包括特征赋权、直觉模糊信息整合，替代指标的排名和未来研究的主要挑战，并陈述了多个领域对这些方法的运用。

4.5 多模数据可视化融合方法

可视化方式是分析多模数据的重要方式，多变量关系展示可以展现单独展示所不能获取的信息。但随着分析变量和节点的增多，可视化效果和信息量难以同时保持平衡。多重网络融合的关键在于如何对由多重网络形成的异构信息网络结构进行重新解构，充分利用共现这一独特视角辅助网络融合。

Leydesdorff[48]把作者-期刊-关键词的特征项联系起来，将不同类型节点在同一网络中进行展现，可以分析同一类型节点间或不同类型节点间的关系，更加真实地反映研究网络。Morris 等[49-50]借助两个共现矩阵相同特征项之间的关联，开发了交叉

图和时间线技术并进行了应用研究，很好地解决了揭示两种特征项关联的可视化问题。为将文本信息融入统计数据，Antal 等[51]分析了文本信息的统计向量的相似性和独立性是否与专家评估信息一致，并且提出了基于文本信息的领域变量的关系抽取应用方法，以支持 Bayesian 网络结构的学习[52]。庞弘燊[53]借鉴 Morris 的交叉图显示方式，并对其做出改进，改进后的交叉图除了可显示两个特征项之间的关联关系之外，还可以显示三个特征项之间的共现关系。魏绪秋等[54]以知识管理主题的研究文献为例，构建作者、关键词和期刊 3 模网络，进行实证分析和可视化，揭示知识管理主题的内在规律及发展趋势。Ghani 等[55]总结了高维数据可视化的方法，并采用分而治之的方法，从设计的流程出发，提出多模社会网络可视化分析方法，并设计了点线平行带的可视化分析效果。van den Elzen 等[56]设计了不受领域专家干涉的通过选择和聚合实现多模数据可是化的方式。史庆伟等[58]和 Xu 等[57]基于时间窗口改进 LDA 主题模型算法，将主题、作者与时间关联构成三维分析分析模型，用于挖掘科技文献数据中不同时间阶段统一主题关注强度的变化情况、内部演化规律及作者兴趣的变化。

对比当前主要的多数据类型关系融合方法，五种多数据类型关系融合方法在融合的机理、方法和复杂程度上存在明显差异，但当前缺少不同方法的对比分析，因此，针对科学计量中的多关系融合方法的适用对象与场景尚不明确。

4.6　数据融合在科学计量中的应用

1）提高文本分析中的主题自动识别的准确度

针对文本分析存在主题词空间维度过高和引文分析中对于引文数据的积累需要较长的时间。同时，在利用标引词进行研究领域描述时，都会面临词语本身所固有的一些问题，如一词多义等。因此，不少研究人员将参考文献和词结合起来用于研究领域分析，词和参考文献的结合有两种不同的方法，一种是利用参考文献作为词间关系的限定，另一种是通过引用构建参考文献-词之间的关系。Calero-Medina 等[59]利用词共现和引文网络分析相结合的方法分析了科学出版物间知识的创造和流动过程。He 等[60-61]提出文本超链接结构、同被引和文本内容相结合的 Web 文本聚类方法。将文本超链接结构作为相似度测算的主要方法，用文本内容相似度来调节超链接的强度，然后通过线性相加方法与同被引整合为一个加权邻接矩阵。Wang 等[62]提出一种用于检索网页的“内容-链接”耦合的聚类算法，将出链、入链、术语三者整合用于提高检索效果。Janssens 等[63-65]将 fisher 逆卡方方法用于文本和文献计量方法的融合，极大的改进了信息融合的准确性，并将方法用于学科分类和主题演化等多个领域。Zhang 等[66]构造了线性权重模型，将基于 IPC 的类别相似分析与基于文本内容分析的语义分析进行专利组合的混合相似度计算。郭红梅等[67]基于多重文本关系图中 clique 子团聚类的主题识别方法，将抽取某主题相关论文集中的术语和术语问的共现、句法和语义关系，构建多重文本关系叠加模型，提高了文本资源中的核心主题识别的效率和准确性。

2）提高科学评价的准确度

多源异质网络的评价增加了更多的评价信息，让评价更加客观。Amjad 等[68]提出了基于主题的异质网络的 TH 排序法（Topic-based Heterogeneous Rank）。TH 排序用三种异构网络分别表示作者和文章之间的关系、文章和期刊之间的关系，在计量任何一种实体的排名时，都参照 Pagerank 的思路，考虑与其他实体的联系，通过矩阵变化将之与引文网络结合，通过综合计算某主题作者、期刊的排名情况。Du 等[69]提出了基于异质网络的发明人重要性排名，辅助识别重要的研发人员，该异质网络由合作网络和专利-发明人网络构成，借助 LDA 主题模型描述发明人的研究主题，结合网络中的同质和异质关系对发明人进行排序，实证证明了该方法适用于专利网络且优于 Pagerank 排名。

5　总结与讨论

科学计量分析的多源数据融合方法多从单一对象或单一关系着手，将多个对象或对象之间多种关系进行融合分析的研究较少，且已有关系融合方法中，简单的线性融合居多，尚未形成系统地面向科技文献主题识别的多关系融合方法。但多源缘于数据融合的复杂性，对于同一个分析目标的主要的数据关系之间存在相关关系，并非相互独立的，因此采用线性运算不能很好地实现数据关系融合。

科学计量中多源数据融合的提高和突破需要坚实的理论基础，缘于数据驱动的科学计量学中多源数据融合需要更复杂深入的计算技术驱动大规模、异构数据分析。科学计量作为以应用为主的研究方法，可以借鉴数理领域的数据融合方法，并根据自

身计量分析的特点，形成科学计量分析的多源数据分析方法。图 6 是本文提出的未来科学计量多源数据融合模式。

第一，前期融合-数据源与数据类型融合。获取更多的数据源，期刊论文、会议信息、学位论文、专利信息、项目信息、著作信息，以及产业经济数据都应纳入计量分析范围。不同的数据类型所反映的科技信息侧重点不同，例如科技论文侧重基础科学研究产出，专利侧重技术创新，而产业经济数据是对科技市场信息的把握。因此，只有全面考虑多源信息，科学计量分析才能得到更客观的分析结果。

未来，一方面需要更多地集成不同来源的多类型科技信息。在数据源方面更多关注新兴商业数据库、新建领域信息共享平台、社交媒体等，在数据类型方面加强产业经济信息、舆情数据等对传统科技文献信息的补充，同时对不同数据源及不同的数据类型进行比对与优选，以达到质量控制的目的。另一方面优化或创新数据结构规范流程与方法，通过智能化、自动化方式实现不同数据的解构、映射以及重组，提高数据集成效率，降低人工成本。

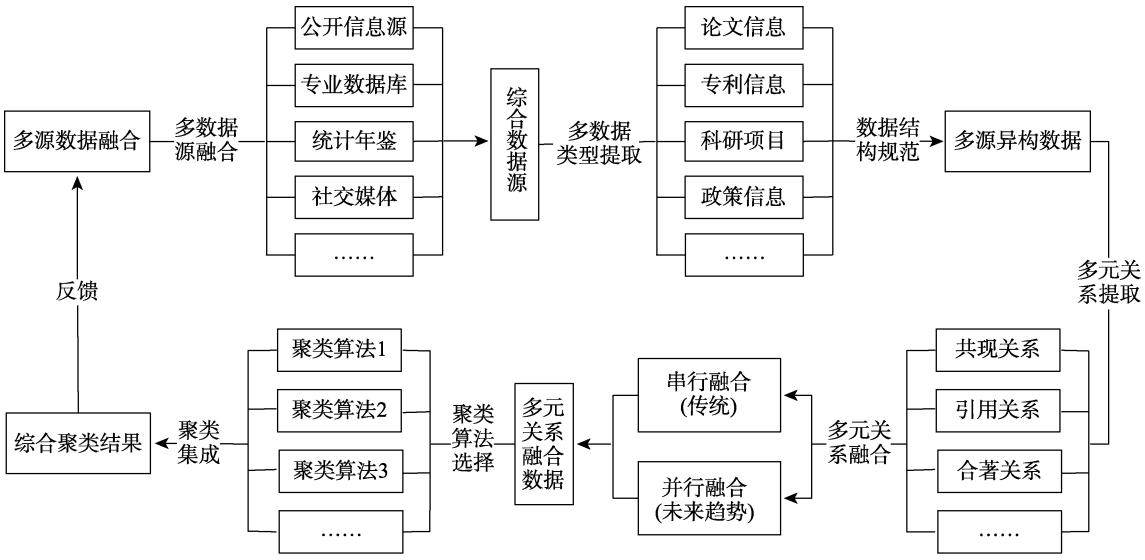

图 6 未来科学计量多源数据融合模式

第二，中期融合-数据关系融合。在获取了多源融合关系的基础上，进一步开展多关系计量分析。因为任意两个计量实体间的关系都具有特定含义，不同关系对研究目标的表征程度不尽相同，理清不同关系的本质特征，并根据特定的研究问题筛选出具有融合价值的多元关系类型，是当前数据关系融合的难题。

未来，在关系复杂、关联实体较多的情况下，可利用超网络、元路径等表示多元关系，在间接关联关系获取时需建立完善的计算关联权重的方法体系，根据路径长短分层次获取这些关系，提升各种关系对目标的表征精度。同时，新的数据分析方法的不断出现，使得未来数据关系融合算法存在很大的改进和提升空间。一方面对其他领域数据关系融合方法在科学计量领域的适用性进行验证，另一方面结合研究目标探索更多的数据关系融合方法，如改变单一的串行融合、并行融合或可视化融合方式，将多种融合方法进行综合集成，或者提高融合知识单元的细粒度，以获取蕴含更大信息量的融合结果。

第三，后期融合-聚类融合。聚类集成研究是机器学习领域的活跃研究热点，从检索到的文献来看，在应用科学计量方法分析科学结构的成果中还没有发现关于聚类集成算法使用的相关研究。优秀的聚类分析方法日益增多，如何选择最为有效的聚类算法和一致性函数，需要引入聚类集成算法将多种聚类结果融合为一个最终的理想结果。

未来，高维数据是需要关注的重点。需要根据具体计量目标与数据特征选择恰当的聚类算法与一致性函数，使集成结果能够从不同角度反映高维数据的综合特性，进一步提升运算效率与聚类结果的质量。同时对聚类集成算法在科学计量领域的应用效果进行实证验证及适应性调整。另外，融合后的评估研究、融合中的复杂度与不确定性等问题也是未来科学计量中多源数据融合方法研究的发展趋势。

参 考 文 献

[1] 许海云, 方曙. 科学计量学的研究主题与发展——基于普赖斯奖得主的扩展作者共现分析[J]. 情报学报, 2013, 32(1): 58-67.

[2] Morris S A, Van der Veer Martens B. Mapping research specialties[J]. Annual Review of Information Science And Technology, 2008, 42(1): 213-295.

[3] Xu H Y, Yue Z H, Wang C, et al. Multi-source data fusion study in scientometrics[J]. Scientometrics, 2017, 111(2): 773-792.

[4] Hai Y X, Chao W, Li J R, et al. Study of multi-source data fusion in topic discovery[M]. Singapore: Springer, 2016.

[5] 化柏林. 情报学三动论探析: 序化论, 转化论与融合论[J]. 情报理论与实践, 2009, 32(11): 21-24, 41.

[6] 化柏林. 多源信息融合方法研究[J]. 情报理论与实践, 2013, 36(11): 16-19.

[7] 化柏林, 李广建. 大数据环境下多源信息融合的理论与应用探讨[J]. 图书情报工作, 2015, 59(16): 5-10.

[8] 曲建升, 刘红煦. 知识发现中异构信息标准化处理研究——以资源环境领域文献为例[J]. 图书情报工作, 2016, 60(6): 84-90.

[9] Small H. A general framework for creating large-scale maps of science in two or three dimensions: The SciViz system[J]. Scientometrics, 1998, 41(1-2): 125-133.

[10] Glenisson P, Glänzel W, Janssens F, et al. Combining full text and bibliometric information in mapping scientific disciplines[J]. Information Processing & Management, 2005, 41(6): 1548-1572.

[11] van Den Besselaar P, Heimeriks G. Mapping research topics using word-reference co-occurrences: A method and an exploratory case study[J]. Scientometrics, 2006, 68(3): 377-393.

[12] Jo Y, Lagoze C, Giles C L. Detecting research topics via the correlation between graphs and texts[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 370-379.

[13] Shibata N, Kajikawa Y, Takeda Y, et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications[J]. Technovation, 2008, 28(11): 758-775.

[14] Wen B, Horlings E, van der Zouwen M, et al. Mapping science through bibliometric triangulation: An experimental approach applied to water research[J]. Journal of the Association for Information Science and Technology, 2017, 68(3): 724-738.

[15] Dong K, Xu H Y, Luo R, et al. An integrated method for interdisciplinary topic identification and prediction: a case study on information science and library science[J]. Scientometrics, 2018, 115(2): 849-868.

[16] Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3: 583-617.

[17] Sun Y, Han J, Yan X, et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.

[18] Sebastian Y, Siew E G, Orimaye S O. Predicting future links between disjoint research areas using heterogeneous bibliographic information network[C]// Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2015: 610-621.

[19] Jensen S, Liu X, Yu Y, et al. Generation of topic evolution trees from heterogeneous bibliographic networks[J]. Journal of Informetrics, 2016, 10(2): 606-621.

[20] Leydesdorff L, Rafols I. Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations[J]. Journal of Informetrics, 2011, 5(1): 87-100.

[21] Latapy M, Magnien C, Del Vecchio N. Basic notions for the analysis of large two-mode networks[J]. Social Networks, 2008, 30(1): 31-48.

[22] 高晨晖, 姜晓睿, 叶政君, 等. 基于异质学术超网的文献评价[J]. 情报学报, 2016, 35(8): 826-837.

[23] Newman M E J. Who is the best connected scientist? A study of scientific coauthorship networks[M]// Ben-Naim E, Frauenfelder H, Toroczkai Z (eds). Complex Networks. Berlin: Springer, 2004: 337-370.

[24] Morris S A, Yen G G. Construction of bipartite and unipartite weighted networks from collections of journal papers[J]. arXiv preprint physics/0503061, 2005.

[25] Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities[J]. Physical Review E, 2009, 80(1): 016105.

[26] Batagelj V, Cerinšek M. On bibliographic networks[J]. Scientometrics, 2013, 96(3): 845-864.

[27] Cerinšek M, Batagelj V. Network analysis of Zentralblatt MATH data[J]. Scientometrics, 2015, 102(1): 977-1001.

[28] 周杰, 刘玉琴, 曾建勋. 学术研究主体与研究内容间的关联关系可视化方法[J]. 现代图书情报技术, 2012(11): 92-97.

[29] Song J, Tang S, Liu X, et al. A modularity-based method reveals mixed modules from chemical-gene heterogeneous network[J]. PLoS ONE, 2015, 10(4): e0125585.

[30] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3): 036106.

[31] Newman M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(23): 8577-8582.

[32] Barber M J, Modularity and community detection in bipartite networks[J]. Physical Review E, 2007, 76(2): 066102.

[33] Comar P M, Tan P N, Jain A K. Simultaneous classification and community detection on heterogeneous network data[J]. Data Mining and Knowledge Discovery, 2012, 25(3): 420-449.

[34] Liu X, Liu W C, Murata T, et al. A framework for community detection in heterogeneous multi-relational networks[J]. Advances in Complex Systems, 2014, 17(6): 1450018.

[35] Nagurney A. Supernetworks: An introduction to the concept and its applications with a specific focus on knowledge supernetworks[D]. Amherst: University of Massachusetts Amherst, 2005.

[36] 王众托. 关于超网络的一点思考[J]. 上海理工大学学报, 2011, 33(3): 229-237.

[37] 胡枫. 复杂超网络的结构、建模及应用研究[D]. 西安: 陕西师范大学, 2014.

[38] 滕立. 基于超网络的作者-机构-国家混合共现网络研究[J]. 情报学报, 2015, 34(1): 28-36.

[39] Lange T, Buhmann J M. Fusion of similarity data in clustering[C]// Proceedings of the Conference on Advances in Neural Information Processing Systems, Vancouver, British Columbia,

Canada, 2005: 723-730.

[40] Wang H, Huang H, Ding C, et al. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization[J]. Journal of Computational Biology, 2013, 20(4): 344-358.

[41] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data[C]// Proceedings of the 28th International Conference on Machine Learning. Madison: Omnipress, 2011: 809-816.

[42] Wang F, Li T, Zhang C S. Semi-supervised clustering via matrix factorization[C]// Proceedings of the SIAM International Conference on Data Mining, Atlanta, Georgia, USA, 2008: 1-12.

[43] Wang H, Huang H, Ding C. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix trifactorization[C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2011: 279-284.

[44] Liu Y, Wang T, Ji X S, et al. Detecting communities in 2-mode networks via fast nonnegative matrix trifactorization[J]. Mathematical Problems in Engineering, 2015, 2015: Article ID 937090.

[45] Dunlavy D M, Kolda T G, Kegelmeyer W P. Multilinear algebra for analyzing data with multiple linkages[M]// Kepner J, Gilbert J (eds). Graph Algorithms in the Language of Linear Algebra, 2011: 85-114.

[46] Snidaro L, García J, Llinas J. Context-based information fusion: a survey and discussion[J]. Information Fusion, 2015, 25: 16-31.

[47] Xu Z, Zhao N. Information fusion for intuitionistic fuzzy decision making: an overview[J]. Information Fusion, 2016, 28: 10-23.

[48] Leydesdorff L. What can heterogeneity add to the scientometric map? Steps towards algorithmic historiography[J]. arXiv preprint arXiv:10020532, 2010.

[49] Morris S, DeYong C, Wu Z, et al. DIVA: a visualization system for exploring document databases for technology forecasting[J]. Computers & Industrial Engineering, 2002, 43(4): 841-862.

[50] Morris S A, Yen G G. Crossmaps: Visualization of overlapping relationships in collections of journal papers[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(suppl 1): 5291-5296.

[51] Antal P, Glenisson P, Fannes G. On the potential of domain literature for clustering and Bayesian network learning[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 405-414.

[52] Antal P, Fannes G, Timmerman D, et al. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors[J]. Artificial Intelligence in Medicine, 2004, 30(3): 257-281.

[53] 庞鸿燊. 基于多重共现的知识发现方法研究[D]. 北京: 中国科学院大学, 2012.

[54] 魏绪秋, 李长玲. 基于作者-年份-关键词网络的科研合作行为研究——以图书情报学为例[J]. 情报杂志, 2014, 33(11): 117-123.

[55] Ghani S, Kwon B C, Lee S, et al. Visual analytics for multimodal social network analysis: A design study with social scientists[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2032-2041.

[56] van den Elzen S, van Wijk J J. Multivariate network exploration and presentation: From detail to overview via selections and aggregations[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 2310-2319.

[57] Xu S, Shi Q W, Qiao X D, et al. Author-Topic over Time (AToT): A dynamic users' interest model[C]// Proceedings of the Conference on Mobile, Ubiquitous, and Intelligent Computing. Berlin: Springer, 2014, 274: 239-245.

[58] 史庆伟, 乔晓东, 徐硕, 等. 作者主题演化模型及其在研究兴趣演化分析中的应用[J]. 情报学报, 2013, 32(9): 912-919.

[59] Calero-Medina C, Noyons E C M. Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field[J]. Journal of Informetrics, 2008, 2(4): 272-279.

[60] He X F, Ding C H Q, Zha H Y, et al. Automatic topic identification using webpage clustering[C]// Proceedings of the 2001 IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2001: 195-202.

[61] He X F, Zha H Y, Ding C H Q, et al. Web document clustering using hyperlink structures[J]. Computational Statistics & Data Analysis, 2002, 41(1): 19-45.

[62] Wang Y T, Kitsuregawa M. Evaluating contents-link coupled web page clustering for web search results[C]// Proceedings of the Eleventh International Conference on Information and Knowledge Management. New York: ACM Press, 2002: 499-506.

[63] Janssens F, Zhang L, De Moor B, et al. Hybrid clustering for validation and improvement of subject-classification schemes[J]. Information Processing & Management, 2009, 45(6): 683-702.

[64] Janssens F, Glänzel W, De Moor B. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 360-369.

[65] Janssens F. Clustering of scientific fields by integrating text mining and bibliometrics[M]. Leuven: Katholieke Universiteit Leuven, 2007.

[66] Zhang Y, Shang L, Huang L, et al. A hybrid similarity measure method for patent portfolio analysis[J]. Journal of Informetrics, 2016, 10(4): 1108-1130.

[67] 郭红梅, 孔贝贝, 张智雄. 基于多重文本关系图中 clique 子团聚类的主题识别方法研究[J]. 情报学报, 2017, 36(5): 433-442.

[68] Amjad T, Ding Y, Daud A, et al. Topic-based heterogeneous rank[J]. Scientometrics, 2015, 104(1): 313-334.

[69] Du Y P, Yao C Q, Li N. Using heterogeneous patent network features to rank and discover influential inventors[J]. Frontiers of Information Technology & Electronic Engineering, 2015, 16(7): 568-578.

[70] Morris S A, Yen G G. Construction of bipartite and unipartite weighted networks from collections of journal papers[J]. Physics, 2005.

（责任编辑 车 尧）